Doctoral Dissertation

Incremental Learning for Large-Scale Stream Data and Its Application to Cybersecurity

大規模ストリームデータのための追加学習とサイバーセキュリティへの応用

July 2015

Graduate School of Engineering

Kobe University

Siti Hajar Aminah Binti Ali

Acknowledgements

First of all, *Alhamdulillah* thanks to our creator, *Allah*, for the continuous blessings and strengths which had been given to me and my beloved family.

I am extremely grateful to my supervisor, Professor Dr. Seiichi Ozawa, for his guidance, support and helpful suggestions throughout my Ph.D study. I wish to thanks Associate Professor Dr. Toshiaki Omori and Dr. Jun Kitazono for willing to give advices and guidances as to improve the quality of my research. Special thanks go to Professor Dr. Masakatu Morii and Professor Dr. Chikara Ohta for the suggestions and comments to improve my dissertation report.

My family deserves special mention for their constant support and their role of being the driving force towards the success of my Ph.D project especially my dearest husband Mohd Fahrul Radzi, my sons: Abdullah Zubair, Abdullah Umar and Abdullah Abbas, and my big family Miskan Siman, Sapiah A. Rahman, Ali Naem, Patimah Ebrahim, Nozli Hanim & Zulkifli, Nor Faraliza & Fuad, Mohd Fahrulnizam & Nuwal, Noor Azila & Rushdi, Mohd Firdaus, Ahmad Talal & KNazwa, Ahmad Omar & Siha, Ahmad Othman & Fairuz and Ahmad Sallehin. Thank you for being patient and always supporting me towards the end.

My sincere appreciation also goes to everyone whom I may not have mentioned above who have helped me directly or indirectly in the completion of this project especially ES5 lab members and MADANI Postgraduates Comittee.

Finally, I would like to express my gratitude to the Ministry of Education (MOE) Malaysia and University Tun Hussein Onn Malaysia (UTHM) for their scholarship award and study leave. This thesis would have not been accomplished without funds which supported by both parties.

Thank you very much.

Abstract

As many human currently depend on technologies to assist with daily tasks, there are more and more applications which have been developed to be fit in one small gadget such as smart phone and tablet. Thus, by carrying this small gadget alone, most of our tasks are able to be settled efficiently and fast. Until the end of 20th century, mobile phones are only used to call and to send short message service (sms). However, in early 21st century, a rapid revolution of communication technology from mobile phone into smart phone has been seen in which the smart phone is equipped by 4G Internet line along with the telephone service provider line. Thus, the users are able to make a phone call, send messages using variety of application such as *Whatsapp* and *Line*, send email, serving websites, accessing maps and handling some daily tasks via online using online banking, online shopping and online meetings via video conferences. In previous years, if there are cases of missing children or missing cars, the victims would rely on the police investigation. But now, as easy as uploading a notification about the loss on *Facebook* and spread the news among *Facebook* users, there are more people are able to help in the search. Despite the advantages that can be obtained using these technologies, there are a group of irresponsible people who take advantage of current technologies for their own self-interest. Among the applications that are usually being used by almost Internet users and also are often misused by cyber criminals are email and websites. Therefore, we take this initiative to make enhancement in cyber security application to avoid the Internet users from being trapped and deceived by the trick of cyber criminals by developing detection system of malicious spam email and Distributed Denial of Services (DDoS)

backscatter.

Imagine that a notice with a logo of Mobile Phone company is received by an email informing that the customer had recently run up a large mobile phone bill. A link regarding the bill is attached for him/her to find out the details. Since, the customer thinks that the billing might be wrong, thus the link is clicked. However, the link is directed to a webpage which displays a status that currently the webpage is under construction. Then the customer closes the page and thinking of to visit the website again at other time. Unfortunately, after a single click actually a malicious file is downloaded and installed without the customer aware of it. That malicious file most probably is a Trojan that capable to steal confidential information from victim's computer. On the next day, when the same person is using the same computer to log in the online banking, all of a sudden find out that his/her money is lost totally. This is one of a worst case scenario of malicious spam email which is usually handled by cybersecurity field. Another different case of cybersecurity is the Distributed Denial of Services (DDoS) attack. Let say, *Company X* is selling flowers via online in which the market is from the local and international customer. The online business of *Company X* is running normally as usual, until a day before mother's day, the webpage of *Company X* is totally down and the prospective customers could not open the webpage to make order to be sent specially for their beloved mother. Thus, the customers would search another company that sells the same item. The *Company X* server is down, most probably because of the DDoS attack where a junk traffic is sent to that company server which makes that server could not serve the request by the legitimate customers. This attack effect not only the profit of the company, but also reputation damage, regular customer turnover and productivity decline.

Unfortunately, it is difficult for a normal user like us to detect malicious spam

email or DDoS attack with naked eyes. It is because recently the spammers and attacker had improved their strategy so that the malicious email and the DDoS packets are hardly able to be differentiated with the normal email and data packets. Once the *Social Engineering* is used by the spammers to create relevant email content in the malicious spam email and when a new campaign of DDoS attack is launched by the attacker, no normal users are capable to distinguish the benign and malicious email or data packets. This is where my Ph.D project comes in handy. My Ph.d is focusing on constructing a detection system of malicious spam email and DDoS attack using a large number of dataset which are obtained by a server that collect double-bounce email and darknet for malicious spam email detection system and DDoS backscatter detection system, respectively. As many up-to-date data are used during the learning, the detection system would become more robust to the latest strategy of the cybercriminal. Therefore, the scenario mentioned above can be avoided by assisting the user with important information at the user-end such as malicious spam email filter or at the server firewall. First of all, the method to learn large-scale stream data must be solved before implementing it in the detection system. Therefore, in Chapter 2, the general learning strategy of large-scale data is introduced to be used in the cybersecurity applications which are discussed in Chapter 3 and Chapter 4, respectively.

One of a critical criterion of the detection system is capable to learn fast because after the learning, the updated information needs to be passed to user to avoid the user from being deceived by the cybercriminal. To process large-scale data sequences, it is important to choose a suitable learning algorithm that is capable to learn in real time. Incremental learning has an ability to process large data in chunk and update the parameters after learning each chunk. Such type of learning keep and update only the minimum information on a classifier model.

Therefore, it requires relatively small memory and short learning time. On the other hand, batch learning is not suitable because it needs to store all training data, which consume a large memory capacity. Due to the limited memory, it is certainly impossible to process online large-scale data sequences using the batch learning. Therefore, the learning of large-scale stream data should be conducted incrementally.

This dissertation contains of five chapters. In Chapter 1, the concept of incremental learning is briefly described and basic theories on Resource Allocating Network (RAN) and conventional data selection method are discussed in this chapter. Besides that, the overview of this dissertation is also elaborated in this chapter. In Chapter 2, we propose a new algorithm based on incremental Radial Basis Function Network (RBFN) to accelerate the learning in stream data. The data sequences are represented as a large chunk size of data given continuously within a short time. In order to learn such data, the learning should be carried out incrementally. Since it is certainly impossible to learn all data in a short period, selecting essential data from a given chunk can shorten the learning time. In our method, we select data that are located in *untrained* or "not well-learned" region and discard data at *trained* or "well-learned" region. These regions are represented by margin flag. Each region is consisted of similar data which are near to each other. To search the similar data, the well-known LSH method proposed by Andoni et al. is used. The LSH method indeed has proven be able to quickly find similar objects in a large database. Moreover, we utilize the LSH ' s properties; hash value and Hash Table to further reduced the processing time. A flag as a criterion to decide whether to choose or not the training data is added in the Hash Table and is updated in each chunk sequence. Whereas, the hash value of RBF bases that is identical with the hash value of the training data is used to select the RBF bases that is near to the training data. The performance results of

the numerical simulation on nine UC Irvine (UCI) Machine Learning Repository datasets indicate that the proposed method can reduce the learning time, while keeping the similar accuracy rate to the conventional method. These results indicate that the proposed method can improve the RAN learning algorithm towards the large-scale stream data processing.

In Chapter 3, we propose a new online system to detect malicious spam emails and to adapt to the changes of malicious URLs in the body of spam emails by updating the system daily. For this purpose, we develop an autonomous system that learns from double-bounce emails collected at a mail server. To adapt to new malicious campaigns, only new types of spam emails are learned by introducing an active learning scheme into a classifier model. Here, we adopt Resource Allocating Network with Locality Sensitive Hashing (RAN-LSH) as a classifier model with data selection. In this data selection, the same or similar spam emails that have already been learned are quickly searched for a hash table using Locally Sensitive Hashing, and such spam emails are discarded without learning. On the other hand, malicious spam emails are sometimes drastically changed along with a new arrival of malicious campaign. In this case, it is not appropriate to classify such spam emails into malicious or benign by a classifier. It should be analyzed by using a more reliable method such as a malware analyzer. In order to find new types of spam emails, an outlier detection mechanism is implemented in RAN-LSH. To analyze email contents, we adopt the Bag-of-Words (BoW) approach and generate feature vectors whose attributes are transformed based on the normalized term frequency-inverse document frequency. To evaluate the developed system, we use a dataset of double-bounce spam emails which are collected from March 1, 2013 to May 10, 2013. In the experiment, we study the effect of introducing the outlier detection in RAN-LSH. As a result, by introducing the outlier detection, we confirm that the detection accuracy is enhanced on

average over the testing period.

In Chapter 4, we propose a fast Distributed Denial of Service (DDoS) backscatter detection system to detect DDoS backscatter from a combination of protocols and ports other than the following two *labeled* packets: Transmission Control Protocol (TCP) Port 80 (80/TCP) and User datagram Protocol (UDP) Port 53 (53/UDP). Usually, it is hard to detect DDoS backscatter from the *unlabeled* packets, where an expert is needed to analyze every packet manually. Since it is a costly approach, we propose a detection system using Resource Allocating Network (RAN) with data selection to select essential data. Using this method, the learning time is shorten, and thus, the DDoS backscatter can be detected fast. This detection system consists of two modules which are pre-processing and classifier. With the former module, the packets information are transformed into 17 feature-vectors. With the latter module, the RAN-LSH classifier is used, where only data located at *untrained* region are selected. The performance of the proposed detection system is evaluated using 9,968 training data from 80/TCP and 53/UDP, whereas 5,933 test data are from *unlabeled* packets which are collected from January 1st, 2013 until January 20th, 2014 at National Institute of Information and Communications Technology (NICT), Japan. The results indicate that detection system can detect the DDoS backscatter from both *labeled* and *unlabeled* packets with high recall and precision rate within a short time.

Finally, in Chapter 5, we discussed the conclusions and the future work of our study: RAN-LSH classifier, malicious spam email detection system and DDoS backscatter detection system.

Contents

List of Figures

List of Tables

List of Abbrevistions

AEB	-	Advanced Entropy-Based
BoW	-	Bag-of-Words
DDoS	-	Distributed Denial of Services
DoS	-	Denial of Services
DNS	-	Domain Name System
FN	-	False Negative
FP	-	False Positive
HTTP	-	Hypertext Transfer Protocol
ID	-	identification
IDF	-	Inverse Document Frequency
IP	-	Internet Protocol
IPCA	-	Incremental Principal Component Analysis
IPv4	-	Internet Protocol version 4
LDOS	-	Low-rate Denial of Services
LSH	-	Locality Sensitive Hashing
MAGIC	-	MAGIC Gamma Telescope
MOE	-	Ministry of Education
NADA	-	Network Anomaly Detection Algorithm
NICT	-	National Institute of Information and Communications Technology
NNs	-	Nearest Neighbors
PCA	-	Principal Component Analysis
PREDICT	-	Protected Repository for Defense of Infrastructure Against Cyber Threats

QR	-	Query/Response flag
RAN	-	Resource Allocating Network
RAN-LSH	-	Resource Allocating Network-Locality Sensitive Hashing
RAN-LTM	-	Resource Allocating Network with Long Term Memory
RAN-MRG	-	RAN with margin-based data selection
RBF	-	Radial Basis Function
RBFN	-	Radial Basis Function Network
SVD	-	Singular Value Decomposition
SVM	-	Support Vector Machines
TCP	-	Transmission Control Protocol
TF	-	Term Frequency
TF-IDF	-	Term Frequency - Inverse Document Frequency
TME	-	Targeted Malicious Email
TN	-	True Negative
TP	-	True Positive
UCI	-	University of California, Irvine
UDP	-	User Datagram Protocol
URL	-	Uniform Resource Locator
UTHM	-	Universiti Tun Hussein Onn Malaysia
53/UDP	-	User Datagram Protocol Port 53
80/TCP	-	Transmission Control Protocol Port 80

List of Symbols

δ	RBF width threshold
σ	RBF width
θ_z	Network output
ε	Error threshold
$\boldsymbol{c}$	Center of RBF
δz	Output margin
Φ	RBF outputs matrix
θ_N	Occurrence frequency threshold
θ_p	Tolerant distance
θ_m	Output margin threshold
θ_a	Accumulation ratio
θ_o	Outlier threshold
θ_z	Highest network output
c^*	Nearest center to $\boldsymbol{x}_i$
$\mathcal{C}$	Index set of RBFs position in hash table
$\boldsymbol{d}$	K-dimensional target vector in which the class label element is one and the remainings are zeros (eg. for class label two with five total different classes; $\boldsymbol{d} = \{0, 1, 0, 0, 0\}$)
d_j^*	LSH distance
d	One of documents in D
$\boldsymbol{D}$	Target matrix
D	Documents in a corpus
E	Error
$f_d(t)$	Frequency of term t in a document d
F^M	Margin flag

F^O	Outlier flag
F^R	Response flag
$\boldsymbol{h}$	Hash value
$\mathcal{H}$	Hash table
$H(v_i)$	Hash function
I_{DB}	Davies-Bouldin index
$N_{d'}$	Number of training data
$N_{d''}$	Number of test data
N_e	Number of occurrence of similar data in each entry e
$N_{\mathcal{H}}$	Number of entries in hash table
N_o	Number of initial training data
P	Number of partitions
$\mathcal{R}$	Index set of selected RBFs
S_i	The variance of data allocated to the ith segment of P partition
t	Spesific term in document d
$\boldsymbol{U}_l$	Eigenvectors
$\boldsymbol{V}$	Projection vectors
v_i^+	Upper values of typical projections v_i
v_i^-	Lower values of typical projections v_i
$\boldsymbol{W}$	Connection weights
$\boldsymbol{x}$	Feature vector
$\bar{\boldsymbol{x}}$	Prototype
$\boldsymbol{X}'$	Training data
$\boldsymbol{X}''$	Test data
$\boldsymbol{X}_o$	Initial training data
$\tilde{\boldsymbol{X}}_s$	All data in each s subsets of unique hash values
$\boldsymbol{y}(\boldsymbol{x})$	RBF outputs
$\boldsymbol{z}(\boldsymbol{x})$	Network outputs

z_{c1}	Largest network output
z_{c2}	Second largest network output

Chapter 1

Introduction

1.1 Background

The rapid growth of storage technology and computer networks has brought the opportunity for researchers to get involved in the processing of large-scale stream data which consists of collecting data in real time, storing, mining and analyzing the knowledge from data. Along with the development of the internet and sensor technologies, a large amount of data is continuously generated in our daily life and such data are typically provided as a time series of large data chunk, so-called *big stream data*. For such a large-scale stream data, it is usually difficult to learn all the data given at the same time or within a short period. Stream data are characterized as an unlimited sequence of data which are given continuously in a short time and whose data distribution is changed over time[1)]. There are at least three major issues to be solved: (1) system that capable to monitor for each second of activity continuously, (2) the growing number of data and (3) how to learn fast. To solve the first issue, it seems that human resource is not an option since human need to rest and could not maintain our focus for a long time, besides our limit to process complicated calculation with high speed. Thus, machine learning is a good alternative to replace human resource which is capable to work non-stop with high speed and with a consistent working performance. Since machine learning is used to solve the first issue, the solution for the other two remaining issues must be related to machine learning too. As for the second issue, it can be solved by learning incrementally using machine learning approach which able to keeps only important information on the previously given data.

Whereas, to cater third issue, the computational time need to be reduced by removing unimportant calculations. The redundant calculations to be discarded should be selected carefully so that the performance would not degrade much. Therefore, in this dissertation, a novel technique to select essential data and to learn Radial Basis Function (RBF) bases locally is introduced which based on Locality Sensitive Hashing (LSH) approach.

Among the research area within the scope of large-scale stream data stream study is cybersecurity application that meets the criteria of stream data. Those criteria are the generation of unlimited data and the characteristics that represent a class label are constantly changing when a cyber criminals attack's techniques evolved. Since more and more people rely on electronic gadgets (i.e., computers, smartphones and tablets) and internet connections, there are those who try to manipulate the advantages of these technologies for *cyberbullying*. Because of this reason, the importance of cyber security research has raised up in order to avoid users being deceived by the recent tactics of cyber criminals. For example, the spammers send spam emails that contain malicious software to steal important data from a particular individual or organization and the attackers send an attack toward certain websites via Distributed Denial of Services (DDoS) to prevent other users from accessing certain website. Both examples of cyber security applications have been developed in this study to detect any harmful input (i.e., malicious spam emails and DDoS backscatter) that would contribute to adverse effect on the users.

Technology without borders has now becoming preference for human to communicate to each other. It is because this technology is not only very fast, but also is easy to be used to facilitate our daily tasks. One of striking example is email application that functions as virtual letter. Although the letter was still used to send original copy of document, usually an email would be sent prior to the original copy. Email is not just used by those registered companies and organizations, in fact email is also used by individuals for personal usage. Due to

the efficiency of email application, most companies and organizations own their internal email system to distribute important information from board of executive to their subordinates without keying in one by one email addresses. It is quite problematical task if the organizations have hundreds of staff as their subordinate. Apart of simplicity and fast operation offered by email application, an email address has also becoming an identification (ID) of Internet users in cyber world. For example, to sign up a social media like *Facebook*, an email address is required as the ID to distinguish every single user of *Facebook*. Since the email application is used by almost all Internet users, thus this application is manipulate by a group of individuals to commit cybercrime. As far as I know, the studies which conducted to combat cybercrime through the medium of email only focused on spam email detection and also malicious emails detection targeted at a specific individual known as Targeted Malicious Email (TME). To the best of my knowledge from previous studies, there are still no studies on the detection of malicious spam email in general, which does not only targeted at TME, but also all email users. Thus, the development of malicious spam email detection system has become another novelty in this dissertation.

Besides the email application, another important application is website or Internet browsing. Currently, the websites are officially being used to represent some organizations by explaining their objectives, vision and general overview. In addition, the websites also usually equipped with online services to assist the deal between the customers and the service provider. For instance, the bank institutions provide the online banking for transferring money using Internet and also online shop for purchasing desired items without having to go to the shop physically. Thus, to attack a specific websites would means that the attacker has intent to bring down the high-profit company and high-profile organizations through cyberattacks or known as DDoS attack. As a consequent of DDoS attack, the website is likely to face server-down in which the customers or Internet users are not able to access the website that needed at specific period. This can

Bibliography

1) Guha, S., Koudas, N., and Shim, K., "Data Streams and Histograms," *ACM Symposium on Theory of Computing*, pp. 471–475, 2001

2) Sun, N., and Guo, Y., "A Modified Incremental Learning Approach for Data Stream Classification," *Sixth International Conference on Internet Computing for Science and Engineering*, Henan, pp. 122–125, 2012

3) Melville, P., and Mooney, R. J., "Diverse Ensembles for Active Learning," *Proc. 21th International Conf. on Machine Learning*, Banff, CA, pp. 584–591, 2004

4) Datar, M., Immorlica, N., Indyk, P., and Mirrokni, V. S., "Locality-Sensitive Hashing Scheme Based on p-stable Distributions," *Proceedings of Symposium on Computational Geometry (SoCG'04)*, pp. 253–262, 2004

5) Andoni, A., and Indyk, P., "Near-Optimal Hashing Algorithms for Approximate Nearest Neighbor in High Dimensions," *Communications of the ACM*, Vol. 51, Issue 1, pp. 117–122, 2008

6) Gu, X., Zhang, Y., Zhang, L., Zhang, D., and Li, J., "An Improved Method of Locality Sensitive Hashing for Indexing Large-Scale and High-Dimensional Features," *Signal Processing*, Vol. 93, Issue 8, pp. 2244–2255, 2013

7) Lee, K. M., and Lee, K. M ., "Similar Pair Identification Using Locality-Sensitive Hashing Technique," *Proceedings of Joint 6th International Conference on Soft Computing and Intelligent Systems (SCIS) and 13th International Symposium on Advanced Intelligent Systems (ISIS)*, pp. 2117–2119, 2012

8) Shen, H., Li, T., Li, Z., and Ching, F., "Locality Sensitive Hashing Based Searching Scheme for a Massive Database," *Proceedings of IEEE SoutheastCon'08*, pp. 123–128, 2008

9) Scheper, C., Cantor, S. and Maughan, D., "PREDICT: A Trusted Framework for Sharing Data for Cyber Security Research," *Proceedings of the First Workshop on Building Analysis Datasets and Gathering Experience Returns for Security (BADGER'11)*, New York, NY, USA, pp. 105–106, 2011

10) The White House Cyberspace, "Cyberspace Policy Review: Assuring a Trusted and Resilient Information and Communications Infrastructure," pp. 1–38, 2009 http://www.whitehouse.gov/assets/documents/Cyberspace_Policy_Review_final.pdf

11) Haykin, S., "Neural Networks: A Comprehensive Foundation," Prentice Hall, 1999

12) Platt J., "A Resource-Allocating Network for Function Interpolation," *Neural Computation*, Vol. 3, Issue 2, pp. 213–225, 1991

13) Kobayashi, M., Zamani, A., Ozawa, S., and Abe, S., "Reducing Computational in Incremental Learning for Feedforward Neural Network with Long-Term Memory," *Proceedings International Joint Conference on Neural Networks*, pp. 1989–1994, 2001

14) Okamoto, K., Ozawa, S., and Abe, S., 'A Fast Incremental Learning Algorithm of RBF Networks with Long-Term Memory," *Proceedings of International Joint Conference on Neural Networks*, pp. 102–107, 2003

15) Ozawa, S., Tabuchi, T., Nakasaka, S., and Roy, A., "An Autonomous Incremental Learning Algorithm for Radial Basis Function Networks," *Journal of Intelligent Learning Systems and Applications*, Vol. 2, pp. 179–189, 2010

16) Slaney, M., and Casey, M., "Locality-Sensitive Hashing for Finding Nearest Neighbors [Lecture Notes]," *Signal Processing Magazine*, IEEE, Vol. 25, Issue 2, pp. 128–131, 2008

17) Ozawa, S., Toh, S. L., Abe, S., Pang, S., and Kasabov, N., "Incremental Learning of Feature Space and Classifier for Face Recognition," *Neural Networks*, Vol. 18, Issue 5-6, pp. 575–584, 2005

18) Ali, S. H. A., Ozawa, S., Nakazato, J., Ban, T., and Shimamura, J., "An Autonomous Online Malicious Spam Email Detection System Using Extended RBF Network," *Proc. Int. Joint Conf. on Neural Networks (IJCNN'2015)*, Killarney, Ireland, 2015

19) Alazab, M., Venkatraman, S., Watters, P. and Alazab, M., "Information Security Governance: The Art of Detecting Hidden Malware," *IT Security Governance Innovations: Theory and Research*, ed: IGI Global, pp. 293–315, 2013

20) Symantec Corporation, "Internet Security Thread Report (ISTR)", pp. 1–119, 2015

21) Karspersky Lab, "Spam and Phishing Statistics Report Q1-2014," 2014

22) Deshmukh, P., Shelar, M. and Kulkarni, N., "Detecting of Targeted Malicious Email," *2014 IEEE Global Conference on Wireless Computing and Networking (GCWCN)*, pp. 199–202, 2014

23) Alazab, M., Layton, R. ,Broadhurst, R. and Bouhours, B., "Malicious Spam Emails Developments and Authorship Attribution," *2013 Fourth Cybercrime and Trustworthy Computing Workshop*, pp. 58–68, 2013

24) Song, J., Inoue, D., Eto, M., Kim, H. C., and Nakao, K., "An Empirical Study of Spam : Analyzing Spam Sending Systems and Malicious Web Servers," *2010 10th Annual International Symposium on Applications and the Internet*, pp. 257–260, 2010

25) Anderson, D. S., Fleizach, C., Savage, S., Voelker, G. M., "Spamscatter: Characterizing Internet Scam Hosting Infrastructure," *Proceedings of the USENIX Security Symposium*, Boston, 2007

26) Xie, Y., Yu, F., Achan, K., Panigrahy, R., Hulten, G., and Osipkov, I., "Spamming Botnets: Signatures and Characteristics," *ACM SIGCOMM Computer Communication Review*, Vol. 38, Issue 4, 2008

27) John, J. P., Moshchuk, A., Gribble, S. D., and Krishnamurthy, A., "Studying

Spamming Botnets Using Botlab," *Proceedings of the 6th USENIX Symposium on Networked Systems Design and Implementation*, pp. 291–306, 2009

28) Chen, Y., Ma, X., and Wu, X., "DDoS Detection Algorithm Based on Preprocessing Network Traffic Predicted Method and Chaos Theory," *IEEE Communications Letters*, Vol. 17, Issue 5, pp. 1052–1054, 2013

29) CERT, "Overview of Attack Trends," http://www.cert.org/archive/pdf/attack_trends.pdf, Apr. 8, 2002

30) Li, L. and Lee, G., "DDoS Attack Detection and Wavelets," *Proceedings of the 12th International Conference on Computer Communications and Networks (ICCCN)*, pp. 421–427, 2003

31) Zhang, J., Qin, Z., Ou, L., Jiang, P., Liu, J. R., and Liu, A. L., "An Advanced Entropy-Based DDoS Detection Scheme," *20IO International Conference on Information, Networking and Automation (ICINA)*, pp. 68–71, 2010

32) Ali, S. H. A., Ozawa, S., Nakazato, J., Ban, T., and Shimamura, J., "An Online Malicious Spam Email Detection System Using Resource Allocating Network with Locality Sensitive Hashing," *Journal of Intelligent Learning Systems and Applications (JILSA)*, Vol. 7, pp. 42–57, 2015

33) Ali, S. H. A., Furutani, N., Ozawa, S., Nakazato, J., Ban, T., and Shimamura, J., "Distributed Denial of Service (DDoS) Backscatter Detection System Using Resource Allocating Network with Data Selection," *Memoirs of Graduate Schools of Engineering and System Informatics, Kobe University*, Vol. 7, pp. 000-000, 2015

34) Attar, V., Sinha, P., and Wankhade, K., "A Fast and Light Classifier for Data Streams," *Evolving Systems*, Vol. 1, Issue 3, pp. 199–207, 2010

35) Polikar, R., Udpa, L., Udpa, S. S., and Honavar, V., "Learn++: An Incremental Learning Algorithms for Supervised Neural Networks," *IEEE Transactions on Systems, Man, and Cybernetics*, Vol. 31, pp. 497–508, 2001

36) Cauwenberghs, G., and Poggio, T., "Incremental and Decremental Support

Vector Machine Learning," *Advances Neural Information Processing Systems (NIPS'2000)*, Vol. 13, 2000

37) Diehl, C. P., and Cauwenberghs, G., "SVM Incremental Learning, Adaptation and Optimization," *Proceedings of the 2003 International Joint Conference on Neural Networks (IJCNN'03)*, pp. 2685–2690, 2003

38) Spüler, M., Rosenstiel, W., Bogdan, M., "Adaptive SVM-Based Classification Increases Performance of a MEG-Based Brain-Computer Interface (BCI)," *Proceedings of the 22nd International Conference on Artificial Neural Networks*, pp. 669–676, 2012

39) Subramanian, K., Suresh, S., and Sundararajan, N., "A Metacognitive Neuro-Fuzzy Inference System (McFIS) for Sequential Classification Problems," *IEEE Transactions on Fuzzy Systems*, Vol. 21, Issue 6, pp. 1080–1095, 2013

40) Lin, H. T., Lin, C. J., and Weng, R. C., "A Note on Platt's Probabilistic Outputs for Support Vector Machines," *Technical Report, Department of Computer Science, National Taiwan University*, 2003

41) Wu, T. F., Lin, C. J., and Weng, R. C., "Probability Estimates for Multi-Class Classification by Pairwise Coupling," *Journal of Machine Learning Research*, Vol. 5, pp. 975–1005, 2004

42) Dan, R., and Kevin, S., "Margin-Bbased Active Learning for Structured Output Spaces," *Proceedings ECML'06*, pp. 413–424, 2006

43) Chellapilla, K., Mityagin, A., and Charles, D. X., "GigaHash: Scalable Minimal Perfect Hashing for Billions of URLs," *Proceedings of the 16th International Conference on World Wide Web*, 2007

44) Bache, K., and Lichman, M., "UCI Machine Learning Repository http://archive.ics.uci.edu/ml Accessed 01 January 2013.

45) Davies, D. L., and Bouldin, D. W., "A Cluster Separation Measure," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Vol. 1, Issue 2, pp. 224–

227, 1979

46) Hsu, C. W., and Lin, C. J., "A comparison of methods for multiclass support vector machines," *IEEE Transactions on Neural Networks*, Vol. 13, Issue 2, pp. 415–425, 2002

47) Ozawa, S., Pang, S., and Kasabov, N., "Incremental Learning of Chunk Data for On-line Pattern Classification Systems," *IEEE Trans. on Neural Networks*, Vol. 19, Issue 6, pp. 1061–1074, 2008

48) Vuong, T. P., and Gan, D., "A Targeted Malicious Email (TME) Attack Tool," *6th International Conference on Cybercrime, Forensics, Education and Training (CFET)*, Christ Church Canterbury, 2012

49) Nagarjuna, B. V. R. R., and Sujatha, V., "An Innovative Approach for Detecting Targeted Malicious E-mail," *International Journal of Application or Innovation in Engineering & Management (IJAIEM)*, Vol. 2, Issue 7, pp. 422–428, 2013

50) Symantec Corporation, "Internet Security Threat Report 2014," Vol. 19, pp. 1–98, 2014

51) Hurcombe, J., "Malicious Links: Spammers Change Malware Delivery Tactics," 2014 http://www.symantec.com/connect/blogs/malicious-links-spammers-change-malware-delivery-tactics

52) Amin, R. M., Ryan, J., and van Dorp, J., "Detecting Targeted Malicious Email Using Persistent Threat and Recipient Oriented Features," *IEEE Security and Privacy Magazine*, Vol. 99, pp. 1–12, 2011

53) Hadnagy, C., "Social Engineering: The Art of Human Hacking," *Wiley*, Indianapolis, 2011

54) Jungsuk, S., "Clustering and Feature Selection Methods for Analyzing Spam Based Attacks," *Journal of the National Institute of Information and Communications Technology*, Vol. 58, Issue 3/4, pp. 35–50, 2011

55) Criddle, L., "What are Bots, Botnets and Zombies?"

http://www.webroot.com/nz/en/home/resources/tips/pc-security/security-what-are-bots-botnets-and-zombies

56) Nazirova, S., "Survey on Spam Filtering Techniques," *Communications and Network*, Vol. 3, pp. 153–160, 2011

57) Ali, S. H. A., Fukase, K., and Ozawa, S., "A Neural Network Model for Large-Scale Stream Data Learning Using Locally Sensitive Hashing," *Neural Information Processing Lecture Notes in Computer Science*, pp. 369–376, 2013

58) Langley, P., "Selection of Relevant Features in Machine Learning," *Proceedings of the AAAI Fall Symposium on Relevance*, LA, pp. 140–144, 1994

59) Oyang Y. J., Hwang, S. C., Ou, Y. Y., Chen, C. Y., and Chen, Z. W., "Data Classification with Radial Basis Function Networks Based on a Novel Kernel Density Estimation Algorithm," *IEEE Trans Neural Networks*, Vol. 16, Issue 1, pp. 225–236, 2005

60) Dai, Y., Tada, S., Ban, T., Nakazato, J., Shimamura, J., and Ozawa, S., "Detecting Malicious Spam Mails: An Online Machine Learning Approach," *Neural Information Processing Lecture Notes in Computer Science*, Vol. 8836, pp. 365–372, 2014

61) Cortes, C. and Vapnik, V., "Support-Vector Networks," *Machine Learning*, Vol. 20, Issue 3, pp. 273–297, 1995

62) Brank, J., Grobelnik, M., Milič-frayling, N., and Mladenič, D., "Feature Selection Using Linear Support Vector Machines," *Proceedings of the 3rd Int. Conf. on Data Mining Methods and Databases for Engineering, Finance, and Other Fields*, pp. 84–89, 2002

63) Robertson, S. E., Walker, S., Jones, S., Hancock-Beaulieu, M. M., and Gatford, M., "Okapi at TREC-3," *Proceedings of the Third NIST Text Retrieval Conference (TREC3)*, Washington, DC, pp. 109–126, 1996 NIST Special Publication 500-225.

64) Olson, P., "The Largest Cyber Attack in History Has Been Hitting Hong

Kong Sites," 2014 http://www.forbes.com/sites/parmyolson/2014/11/20/the-largest-cyber-attack-in-history-has-been-hitting-hong-kong-sites/

65) Lee, D., "Global Internet Slows After 'Biggest Attack in History'," 2013 http://www.bbc.com/news/technology-21954636,

66) Graham-Cumming, J., "The Wednesday Witching Hour: CloudFlare DoS Statistics," 2012 https://blog.cloudflare.com/the-wednesday-witching-hour-cloudflare-dos-st/

67) Jackson, D., "Understanding and Combining DDoS Attacks: A Threat Analysis," 2011 http://www.secureworks.com/assets/pdf-store/articles/Understanding_and_Combating_DDoS_Attacks.pdf

68) Xu, X., Wei, D., and Zhang, Y., "Improved Detection Approach for Distributed Denial of Service Attack Based on SVM," *Third Pacific-Asia Conference on Circuits, Communications and System (PACCS'11)*, pp. 1–3, 2011

69) Furutani, N., Ban, T., Nakazato, J., Shimamura, J., Kitazono, J., and Ozawa, S., "Detection od DDoS Backscatter Based on Traffic Features of Darknet TCP Packets," *Ninth Asia Joint Conference on Information Security (ASIA JCIS)*, pp. 39–43, 2014

70) Xu, X., Wei, D., and Zhang, Y., "Improved Detection Approach for Distributed Denial of Service Attack Based on SVM," *2011 Third Pacific-Asia Conference on Circuits, Communications and System (PACCS)*, pp. 1–3, 2011

71) Saied, A., Overill, R. E., and Radzik, T., "Artificial Neural Networks in the Detection of Known and Unknown DDoS Attacks: Proof-of-Concept," *Proceedings of PAAMS 2014 International Workshops*, Salamanca, Spain, pp. 309–320, 2014

72) H. Hishinuma, "Information Security by NICT and the Government of Japan," 2011 http://www.bic-trust.eu/files/2011/12/slides13.pdf

73) Broomhead, D. S., and Lowe, D., "Radial Basis Functions, Multi-variable Functional Interpolation and Adaptive Networks (Technical Report)," *Royal*

Signal and Radar Establishment (RSRE), Vol. 4148, pp. 1–39, 1988

74) Poggio, T. and Girosi, F, "Networks for Approximation and Learning," *IEEE Trans. on Neural Networks*, Vol. 78, Issue 9, pp. 1481–1497, 1990